

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**



**QUIZ 1: MỐI QUAN HỆ TRONG DỮ LIỆU**

**Học phần: Trắc quan hóa dữ liệu**

**Lớp học phần 19\_21**

**Giảng viên Lý thuyết Bùi Tiến Lên**

**Giảng viên Thực hành Lê Ngọc Thành**

**Các sinh viên thực hiện**

19120467 Ngô Hữu Đang

19120615 Hùng Ngọc Phát

19120684 Hồ Minh Quân

**Thành phố Hồ Chí Minh**

**Tháng 05 năm 2022**

# MỤC LỤC

<b>I. Phân công và đánh giá</b>	<b>2</b>
<b>II. Thu thập dữ liệu</b>	<b>4</b>
<b>III. Khám phá và tiền xử lý dữ liệu</b>	<b>4</b>
1. Dữ liệu có bao nhiêu dòng, bao nhiêu cột ?	4
2. Các dòng có ý nghĩa gì ? Có vấn đề các dòng có ý nghĩa khác nhau không ?	4
3. Dữ liệu có bị lặp không ?	4
4. Ý nghĩa của từng cột	4
5. Kiểu dữ liệu của từng cột	5
6. Tỷ lệ missing của các cột	6
<b>IV. Phân tích và trực quan hóa</b>	<b>7</b>
1. Phân bố dữ liệu của một số cột numerical	7
2. Phân bố của các hàng null	8
3. Tỷ lệ tình trạng bệnh nhân trên toàn cầu	9
4. Số lượng và tỷ trọng quốc gia ở các châu lục	10
5. Phân bố tình trạng bệnh nhân ở các châu lục	10
6. Khảo sát các mối quan hệ trên các cột numeric	13
7. Các mối quan hệ trên 2 biến numerical	13
a. TotalDeaths ~ Population	14
b. TotalRecovered ~ TotalTests	15
c. TotalDeaths ~ TotalCases	17
d. ActiveCases ~ TotalCases	17
8. Mối quan hệ giữa các cột NewCases, NewDeaths, NewRecovered sau khi xử lý giá trị null	18
9. Các mối quan hệ trên nhiều hơn 2 biến numerical	21
a. NewDeaths ~ TotalDeaths + Population	21
b. TotalDeaths ~ TotalCases + TotalTests	22
c. NewCases ~ NewRecovered + ActiveCases	22
<b>V. Tài liệu tham khảo</b>	<b>23</b>

## I. Phân công và đánh giá

MSSV	Tên	Nhiệm vụ
19120467	Ngô Hữu Đang	Tiền xử lý dữ liệu, giải thích data, vẽ các biểu đồ phân bố tình trạng bệnh nhân, phân bố dân số/quốc gia trên toàn cầu, giữa các cột bị null cao.
19120615	Hùng Ngọc Phát	Tiền xử lý dữ liệu, phân tích phân bố các hàng null và vẽ histogram, boxplot, heatmap; xử lý các mối quan hệ trên 3 biến.
19120626	Hồ Minh Quân	Thu thập dữ liệu, giải thích các mối quan hệ trên 2 biến.

Yêu cầu	Tự đánh giá	
Thu thập và tiền xử lý	Nhóm viết được crawler, thu thập được data của nhiều ngày (nhưng chỉ sử dụng ngày 20/04 21/04 để đơn giản). Nhóm tạm thời xử lý được các trường dữ liệu bị NA.	5/5
Chọn lựa, giải thích, trực quan các trường và mối quan hệ giữa chúng	Nhóm đã sử dụng nhiều loại biểu đồ khác nhau trong số các biểu đồ đã được học: scatter plot, bar chart, stacked bar chart, pie chart, histogram, box plot, kernel density estimation plot, joint plot, correlation matrix, heat map. Nhóm có sử dụng thang đo phi tuyến (non-linear scale) trong các biểu đồ. Nhóm xây dựng được các mô hình hồi quy OLS có nhiều biến, vẽ được đường hồi quy, khoảng tin cậy và giải thích được các kết quả phân tích thống kê. Nhóm còn thiếu radar chart và line chart vì không có dữ liệu (feature) phù hợp để trực quan. Nhóm chưa sử dụng các biểu đồ dạng bản đồ, có lẽ để chờ các lab khi được sử dụng <i>màu sắc</i> nhóm sẽ tiến hành trực quan.	45/50
Rút ra ý nghĩa hợp lý sau mỗi dữ liệu được trực	Nhóm nhận xét và liên kết được ý nghĩa giữa nhiều loại biểu đồ. Nhóm đọc và hiểu được các kết quả phân tích thống kê	18/20

quan	(gồm p-value, $R^2$ , ...). Nhóm có kiểm chứng giả thuyết bằng các bài báo từ bên ngoài.	
Xem xét trên nhiều quan hệ, nhiều góc nhìn khác nhau	Cùng một mối quan hệ, nhóm sử dụng 1 hoặc nhiều loại biểu đồ để trực quan, vì mỗi loại biểu đồ làm mất đi một lượng thông tin nhất định. Nhóm có so sánh giữa số liệu trước và sau khi chuẩn hóa để có một cái nhìn khác hơn về đơn vị của dữ liệu.	8/10
Báo cáo bố cục hợp lý	Nội dung được trình bày rõ ràng, chi tiết.	15/15
<b>Tổng (tự đánh giá)</b>		90%

## II. Thu thập dữ liệu

Dữ liệu được thu thập từ trang [www.worldometers.info](http://www.worldometers.info), ban đầu nhóm thử sử dụng Excel để lưu dữ liệu về, tuy nhiên dữ liệu thu thập được không chuẩn nên nhóm đã thay đổi phương thức get data bằng việc dùng Python 3 với các thư viện request, beautifulsoup, được đính kèm trong thư mục source với tên *crawl.py*.

Nhóm đã thu thập dữ liệu trong nhiều ngày, nhưng để đơn giản thì nhóm chỉ sử dụng data của 2 ngày 20/4 và 21/4, trong đó data của ngày 21/4 đóng vai trò chính.

### Hướng dẫn thực thi:

1. Mở Terminal tại thư mục chứa file source code, sau đó nhập lệnh `python crawl.py`. Lưu ý, bảo đảm kết nối Internet đến trang web trước khi chạy.
2. Nhập 1 nếu muốn lấy dữ liệu của 1 ngày trước, 2 nếu muốn lấy dữ liệu của 2 ngày trước
3. Nếu lấy thành công, dữ liệu sẽ được in lên màn hình và lưu xuống file.

## III. Khám phá và tiền xử lý dữ liệu

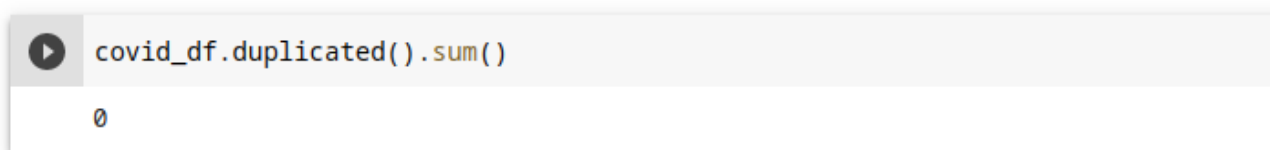
### 1. Dữ liệu có bao nhiêu dòng, bao nhiêu cột ?

- Dữ liệu có 228 dòng và 21 cột

### 2. Các dòng có ý nghĩa gì ? Có vấn đề các dòng có ý nghĩa khác nhau không ?

- Quan sát sơ bộ dữ liệu ta thấy mỗi dòng chứa các thông tin về một quốc gia
- Có vẻ như không có vấn đề, các dòng có cùng ý nghĩa với nhau.

### 3. Dữ liệu có bị lặp không ?



```
covid_df.duplicated().sum()
0
```

- Sau khi kiểm tra thì nhóm thấy dữ liệu không bị lặp

### 4. Ý nghĩa của từng cột

- **Country,Other**: Quốc gia
- **TotalCases**: Tổng số ca nhiễm
- **NewCases**: Số ca nhiễm mới
- **TotalDeaths**: Tổng số ca tử vong
- **NewDeaths**: Số ca tử vong mới
- **TotalRecovered**: Tổng số ca khỏi bệnh
- **NewRecovered**: Số ca khỏi bệnh mới

- **ActiveCases**: Số ca đang điều trị
- **Serious,Critical**: Số ca nguy kịch
- **TotalCases/1M pop**: Tổng số ca nhiễm trên 1 triệu dân
- **Deaths/1M pop**: Tổng số ca tử vong trên 1 triệu dân
- **TotalTests**: Tổng số lần xét nghiệm Covid
- **Test/1M pop**: Tổng số lần xét nghiệm trên 1 triệu dân
- **Population**: Dân số
- **Continent**: Châu lục
- **1 Caseevery X ppl**: X người sẽ có 1 ca nhiễm
- **1 Deathevery X ppl**: X người sẽ có 1 ca tử vong
- **1 Testevery X ppl**: X người sẽ có 1 người đã được xét nghiệm
- **New Cases/1M pop**: Số ca nhiễm mới trên 1 triệu dân
- **New Deaths/1M pop**: Số ca tử vong mới trên 1 triệu dân
- **Active Cases/1M pop**: Số ca đang điều trị trên 1 triệu dân

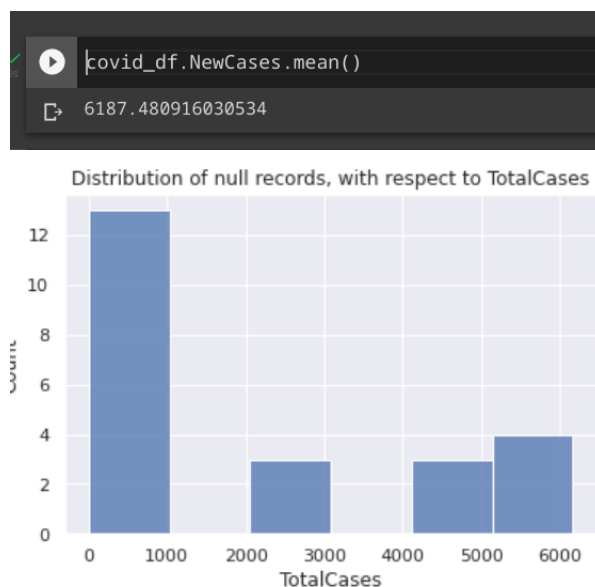
## 5. Kiểu dữ liệu của từng cột

- Các cột loại categorical: *'Country,Other'*, *'Continent'* có kiểu dữ liệu là object(str)
- Các cột loại numerical : *'TotalCases'*, *'NewCases'*, *'TotalDeaths'*, *'NewDeaths'*, *'TotalRecovered'*, *'NewRecovered'*, *'ActiveCases'*, *'Serious,Critical'*, *'TotalCases/1M pop'*, *'Deaths/1M pop'*, *'TotalTests'*, *'Test/1M pop'*, *'Population'*, *'1 Caseevery X ppl'*, *'1 Deathevery X ppl'*, *'1 Testevery X ppl'*, *'New Cases/1M pop'*, *'New Deaths/1M pop'*, *'Active Cases/1M pop'* có kiểu dữ liệu là float64 (int64 + nan).

## 6. Tỷ lệ missing của các cột

	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical		
missing_ratio	0.0	42.5	3.9	69.7	6.6	51.3	6.6	33.8		
count	228.0	131.0	219.0	69.0	213.0	111.0	213.0	151.0		
mean	2226556.8	6187.5	28474.8	47.7	2106254.6	8508.2	91877.8	277.0		
std	7383916.7	19882.3	98819.9	102.2	7289109.2	29567.5	316612.9	874.1		
min	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0		
25%	22581.5	28.5	222.5	1.0	13872.0	22.0	212.0	5.0		
50%	163079.0	181.0	2201.0	10.0	131100.0	245.0	2650.0	20.0		
75%	1063337.5	1355.5	13932.0	41.0	959289.0	2155.5	30915.0	179.0		
max	82553058.0	139849.0	1017609.0	646.0	80355389.0	243700.0	2928053.0	8318.0		
TotalCases/1M pop	Deaths/1M pop	TotalTests	Test/1M pop	Population	Caseevery X ppl	Deathevery X ppl	Testevery X ppl	New Cases/1M pop	New Deaths/1M pop	Ca
0.9	4.8	7.0	7.0	9.000000e-01	0.9	4.8	7.0	42.5	69.7	
226.0	217.0	212.0	212.0	2.260000e+02	226.0	217.0	212.0	131.0	69.0	
141961.9	1150.3	29456351.0	1904397.9	3.493437e+07	1039.9	13765.0	12.1	510.2	2.1	
150973.9	1206.5	106512884.7	3263539.8	1.389877e+08	8837.8	53940.2	29.1	1245.8	3.5	
9.0	2.0	5117.0	5099.0	8.050000e+02	1.0	159.0	0.0	0.0	0.0	
11257.0	172.0	347296.8	164740.8	5.599682e+05	4.0	542.0	0.0	2.5	0.2	
88841.0	765.0	2120364.5	772995.5	5.795650e+06	11.0	1308.0	1.0	56.0	0.7	
233495.5	1844.0	12237504.0	2212139.0	2.184432e+07	89.0	5799.0	6.0	402.0	2.0	
704474.0	6293.0	999047824.0	21807666.0	1.439324e+09	117193.0	623755.0	196.0	9159.0	22.0	5

Các cột NewCases, NewDeaths và NewRecovered có tỷ lệ missing rất cao, nhưng không thể fill các cột này bằng mean, lý do là vì các cột này có mean khá lớn (6187), trong khi trong các quốc gia bị null thì có các quốc gia có tổng số ca (TotalCases) nhỏ hơn 6000 (biểu đồ bên dưới).



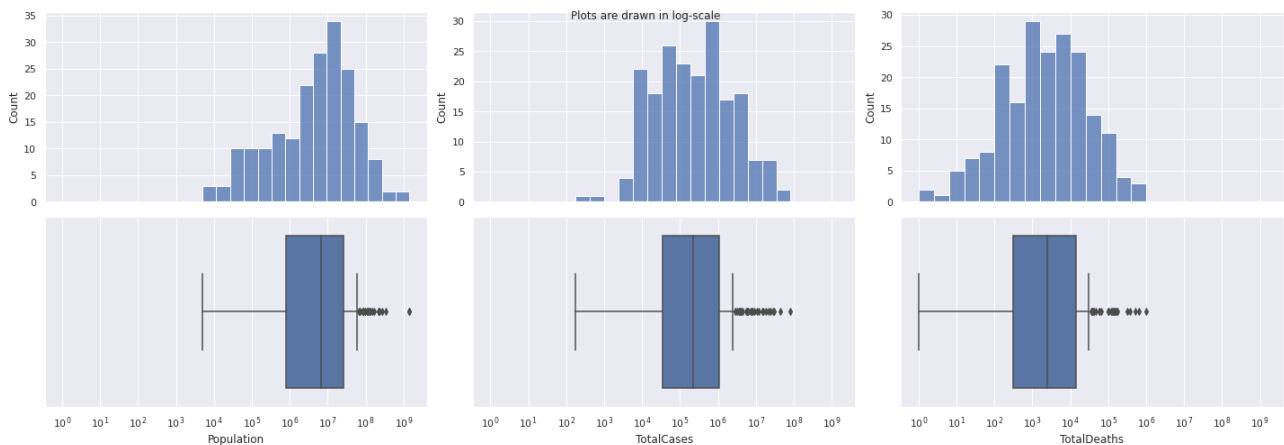
Nhóm có sử dụng data của ngày 20/04 để fill cho ngày 21/04, nhưng kết quả cũng không mấy khả quan. Do đó, giải pháp tạm thời của nhóm trong lab này: drop các cột có tỷ lệ null cao đi. Ta sẽ xem xét lý do các cột này bị null ở bên dưới.

## IV. Phân tích và trực quan hóa

### 1. Phân bố dữ liệu của một số cột numerical

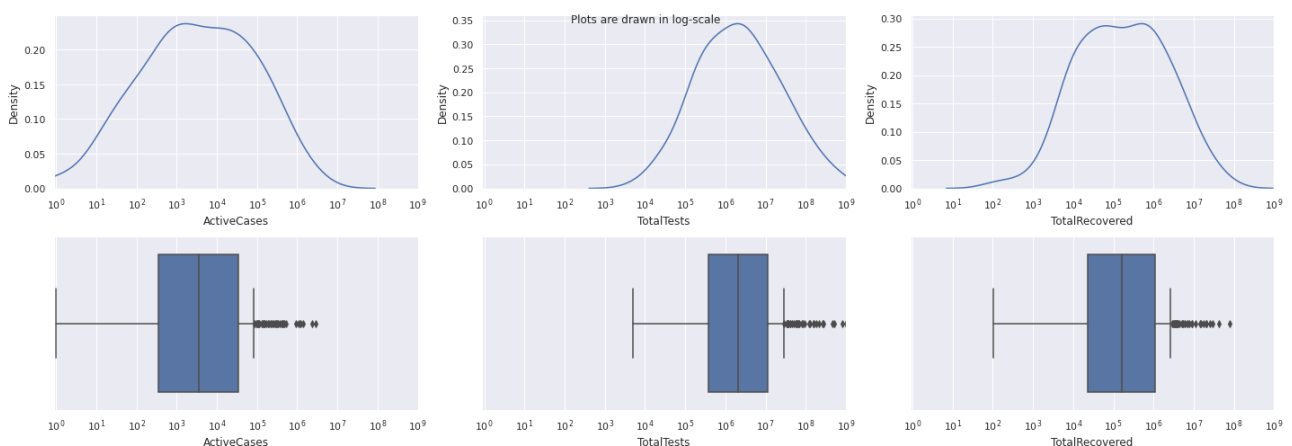
Phân bố dữ liệu của các cột *Population*, *TotalCases*, *TotalDeaths* với biểu đồ *histogram* và *boxplot*.

Lý do phải sử dụng cả 2 là vì histogram không thể hiện được các outlier trong dữ liệu, còn boxplot nếu đứng riêng thì hơi khó hình dung về phân bố.



- Về dân số: dân số của các nước trên thế giới phân bố không đều: 50% các nước có dân số từ khoảng 5000 đến dưới 10 triệu người, 50% còn lại có dân số trên 10 triệu người. Có 1 điểm dữ liệu đặc biệt, với giá trị lên đến hơn 1 tỷ.
- 50% các nước có tổng số ca nhiễm nằm từ khoảng 12,000 đến 1 triệu ca (IQR). Có nước cá biệt nhiễm lên đến khoảng 100 triệu ca. Phân phối bị lệch về bên trái.
- 50% các nước có tổng số ca tử vong nằm trong khoảng 200 đến 11,000 ca (IQR). Có nước cá biệt chết đến trên 1 triệu người. Phân phối bị lệch trái.

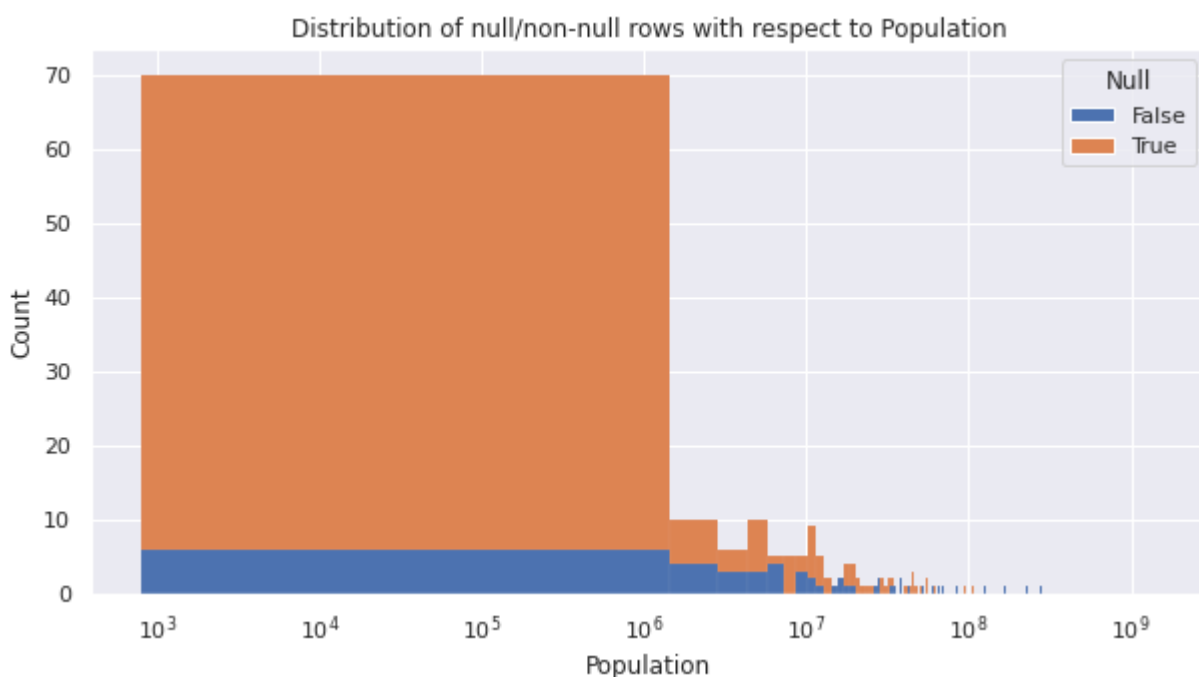
Phân bố dữ liệu của các cột *ActiveCases*, *TotalTests*, *TotalRecovered* với biểu đồ *KDE plot* và *boxplot* với lý do tương tự như trên.





- 50% các nước có số ca đang điều trị từ khoảng 200 đến khoảng 15000 ca (IQR). So với tổng số ca đã nhiễm thì con số này không nhiều, cũng dễ hiểu vì dịch đã xuất hiện được hơn 2 năm.
- 50% các nước đã thực hiện test trên khoảng 500k đến trên 10 triệu người.
- 50% các nước có tổng số ca hồi phục nằm trong khoảng 11,000 đến 1 triệu (IQR). Khoảng này gần match với khoảng tổng số ca nhiễm. Một số ít nước có số liệu khá tích cực, đã hồi phục được gần 100 triệu người.

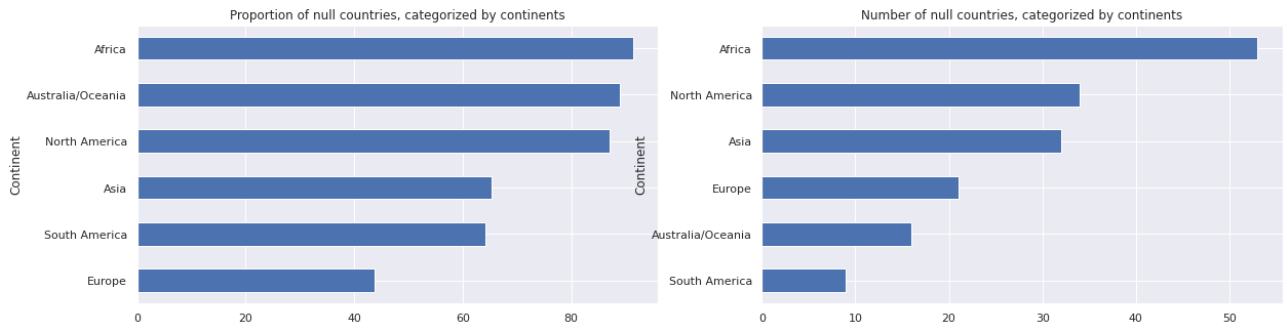
## 2. Phân bố của các hàng null



Nhóm đã vẽ *histogram* với 1000 bins với trục hoành là dân số của các nước, và trục tung là số lượng nước thuộc vào bin đó có ít nhất một trong các cột *NewCases*, *NewDeaths*, *NewRecovered* bị null, và cũng như các nước thuộc trường hợp ngược lại. Để tiện lợi, nhóm sẽ gọi các nước có 3 cột này bị null là “các nước bị null”.

Nhóm cũng sử dụng **log-scale** cho trục  $x$ , vì dữ liệu dân số bị skew khá nhiều nên trục linear-scale sẽ rất khó nhìn. Để đơn giản, ta giả sử các hàng (các nước) bị *null* đơn giản vì việc thu thập số liệu của họ không hiệu quả, không phải do giá trị *null* có ý nghĩa đặc biệt nào khác.

Kết quả trực quan cho thấy phần lớn các nước bị null là những nước có dân số từ 1 triệu trở xuống, tạm gọi là các “nước nhỏ”. Những nước này thu thập số liệu không tốt bằng các “nước lớn”. Tiếp theo ta sẽ xem xét phân bố về mặt địa lý của các nước bị null cao này.

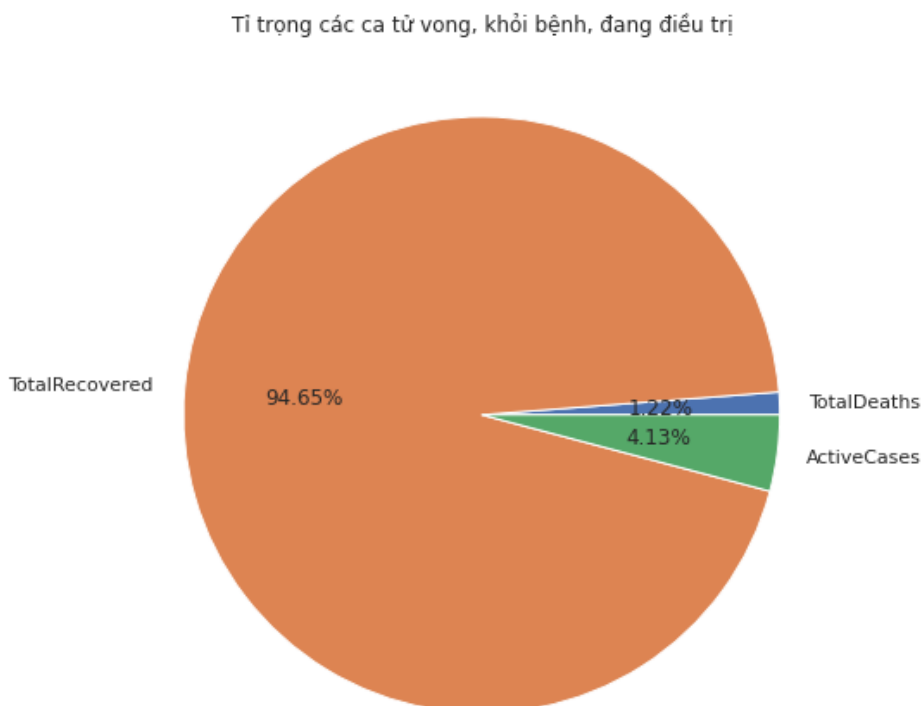


2 biểu đồ lần lượt này thể hiện cho tỉ lệ, và số lượng các nước bị null ở từng châu lục. Kết quả cho thấy châu Phi và châu Úc chiếm phần lớn **tỉ lệ** nước bị null, trong khi châu Âu ít nhất.

Tuy nhiên, khi nhìn sang biểu đồ về **số lượng**, ta lại thấy châu Úc lại đứng gần cuối. Lý do có lẽ châu Úc có ít quốc gia (ta sẽ xem xét điều này sau). Tuy số liệu cho các châu lục khác có thay đổi giữa 2 biểu đồ, châu Phi vẫn đứng đầu ở cả 2, ta có thể “tự tin” kết luận **các nước châu Phi thu thập số liệu kém nhất**. Châu Âu đứng ở thứ hạng thấp ở cả 2 biểu đồ, nên có lẽ **các nước châu Âu thu thập số liệu tốt nhất**.

### 3. Tỉ lệ tình trạng bệnh nhân trên toàn cầu

Trực quan hóa tỉ trọng các ca tử vong, khỏi bệnh, đang điều trị bằng *pie chart* (dùng pie chart để biết được thành phần phần trăm tình trạng của các bệnh nhân)

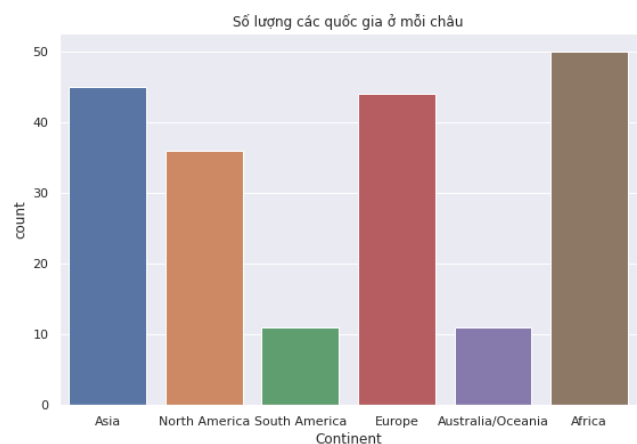
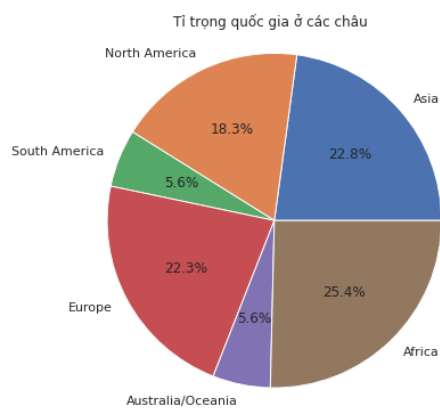


Tổng số ca tử vong chiếm tỉ trọng rất thấp, chỉ 1.22%. Nguyên nhân có lẽ là do đã có vaccine Covid-19 nên số ca tử vong đã giảm đáng kể.

Tổng số ca khỏi bệnh chiếm tỷ trọng rất cao 94.65%, cho thấy căn bệnh này không còn quá nguy hiểm đối với mọi người.

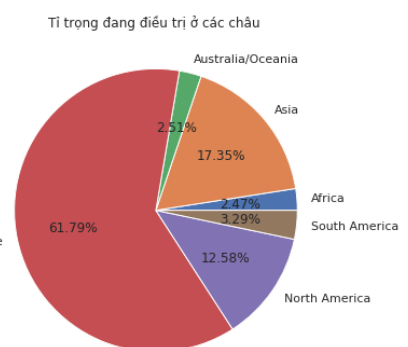
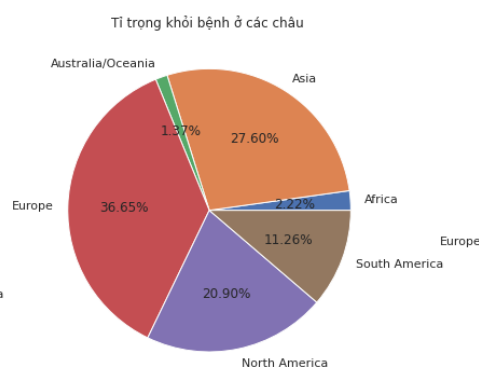
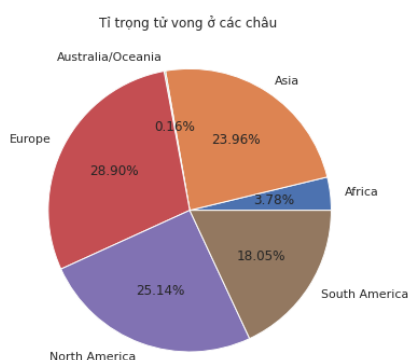
#### 4. Số lượng và tỷ trọng quốc gia ở các châu lục

- Dùng pie chart để thể hiện thành phần phần trăm số lượng các quốc gia của từng châu trong dữ liệu
- Dùng bar chart với các cột được ánh xạ từ pie chart để thể hiện số lượng các quốc gia của từng châu



Từ biểu đồ ta thấy được dữ liệu thu thập từ hầu hết các quốc gia trên thế giới, nhiều nhất là Châu Phi với 50 quốc gia, tiếp sau đó là Châu Á và Châu Âu với 45 và 44 quốc gia.

#### 5. Phân bố tình trạng bệnh nhân ở các châu lục



Các biểu đồ trên thể hiện **tỉ trọng của các châu lục tương ứng với từng tình trạng bệnh nhân**. (nhóm dùng pie chart để có thể dễ dàng so sánh tình trạng của các bệnh nhân ở từng châu)

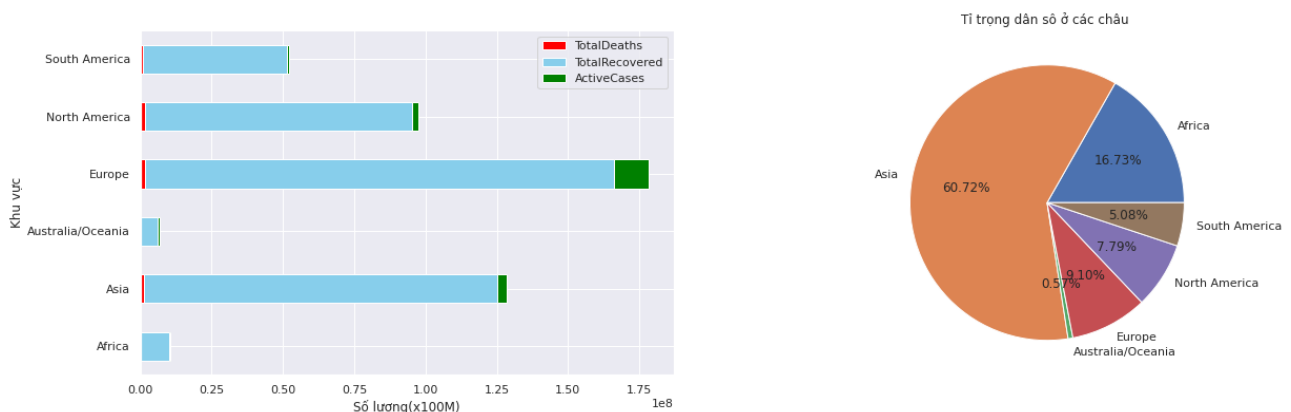
Ta thấy được châu Âu chiếm tỉ trọng cao nhất ở 3 biểu đồ lần lượt là 28.9%, 36.65%, 61.79%, đặc biệt là số ca đang điều trị ở châu Âu chênh lệch rất nhiều so với các châu lục khác hơn 44.44% so với vị trí thứ 2 là châu Á (17.35%).

Bên cạnh đó, châu Đại Dương chiếm tỉ trọng nhỏ nhất ở tỉ trọng tử vong và tỉ trọng khỏi bệnh lần lượt là 0.16%, 1.37%.

Châu Phi có tỉ trọng đang điều trị thấp nhất chỉ với 2.47%

**Dưới góc nhìn ngược lại, ta thử quan sát tỉ lệ phân bố tình trạng bệnh nhân giữa các châu lục bằng stacked bar chart**

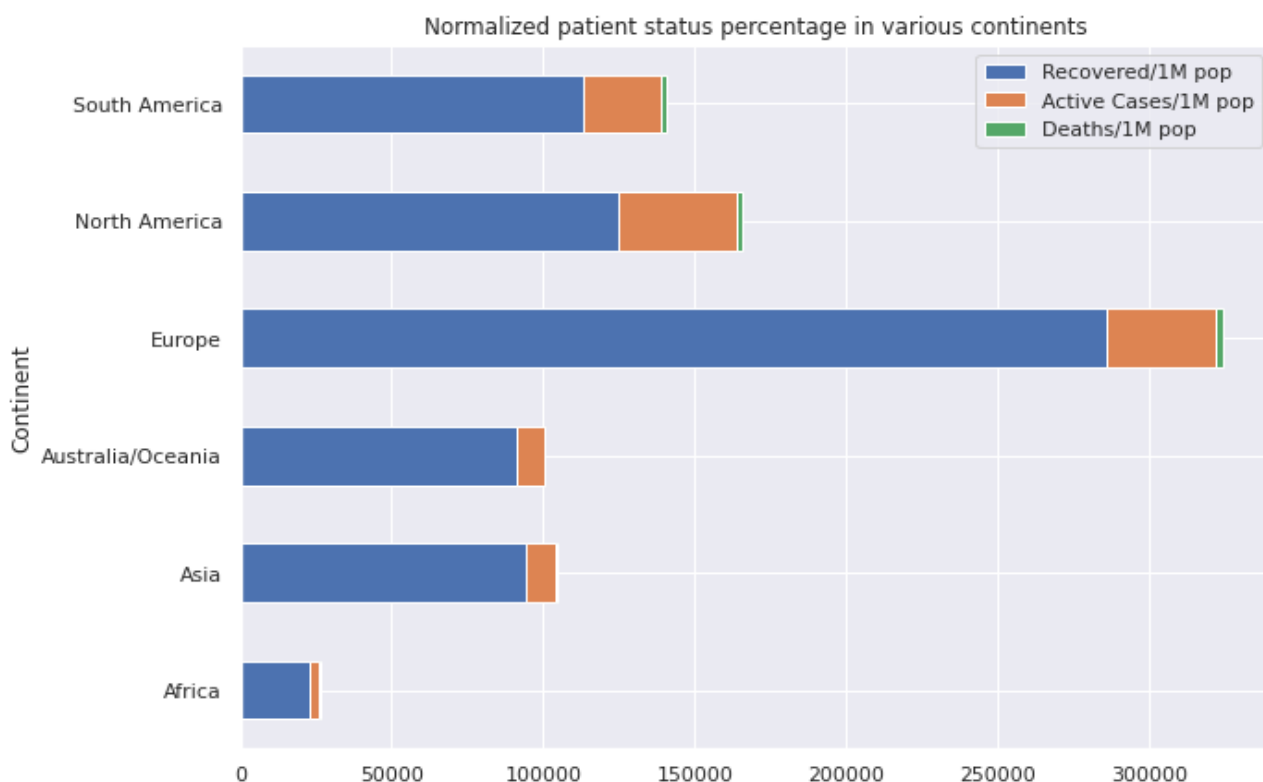
(nhóm dùng thêm pie chart thể hiện tỉ trọng dân số của các châu giúp nhìn thấy được điều bất thường trong bộ dữ liệu)



Giống như các phân tích ở trên, châu Âu có số lượng ca nhiễm covid nhiều nhất với hơn 177 triệu ca nhiễm nhưng dân số chỉ chiếm có 9.1% so với tổng dân số trong dữ liệu (hơn 689 triệu dân).

Điều bất thường tiếp theo là dân số của châu Phi có tỉ trọng cao thứ 2 với 16.73% (hơn 1,2 tỷ dân) lại có số ca nhiễm khá thấp (khoảng 10,6 triệu ca nhiễm).

**Vì dân số các châu lục là khác nhau, ta có thể nhìn thấy điều này rõ hơn thông qua các cột đã được chuẩn hóa:**



Biểu đồ trên thể hiện phân bố tình trạng bệnh nhân của các châu lục, nhưng được thể hiện dưới đơn vị số ca trên 1 triệu dân thay vì số ca.

Sau khi chuẩn hóa (bằng cách lấy tỉ lệ theo phần triệu), châu Âu vẫn là châu lục có tỉ lệ nhiễm cao nhất. Châu Phi là châu lục có tỉ lệ nhiễm thấp nhất.

Dữ liệu trên của châu Phi vẫn không đáng tin vì các dòng của các nước ở châu Phi có nhiều giá trị null nhất. Giả thiết được đưa ra là do không phát hiện ca nhiễm mới hoặc chính phủ ở các nước này không thống kê hết các ca nhiễm.

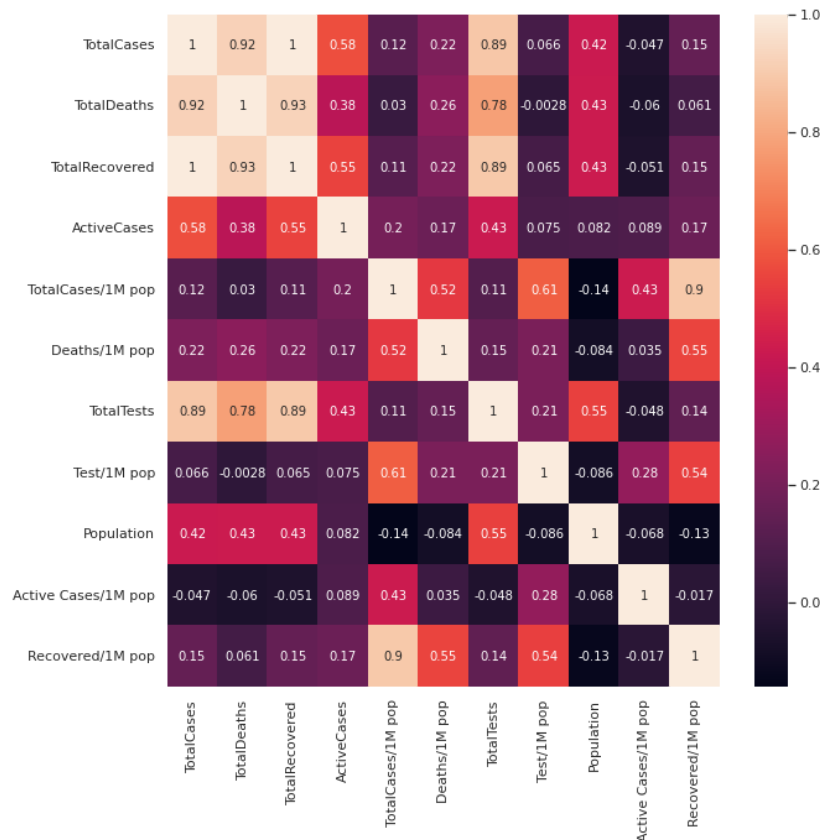
Sau khi lên mạng tìm hiểu thì nhóm sẽ nghiên cứu về giả thiết thứ 2 hơn. Nhóm xin trích lại 1 đoạn trong [bài báo của VNExpress](#).

*"Một dự án nghiên cứu Đại học Njala phát hiện 78% dân số nước này có kháng thể với virus. Tuy nhiên, Sierra Leone chỉ ghi nhận 125 ca tử vong vì Covid-19 kể từ đầu đại dịch.*

*Hầu hết người dân qua đời tại nhà, thay vì bệnh viện. Nhiều ca tử vong không được báo cáo với chính phủ. Tình trạng này phổ biến ở khu vực châu Phi cận Sahara. Cuộc khảo sát gần đây của Ủy*

ban Kinh tế Liên Hợp Quốc về châu Phi cho thấy các hệ thống chỉ ghi nhận khoảng một phần ba số ca tử vong thực tế."

## 6. Khảo sát các mối quan hệ trên các cột numeric



Từ heatmap, có thể thấy các biến sau có tương quan mạnh với nhau (chỉ liệt kê một vài biến):

- TotalCases, TotalDeaths
- TotalCases, TotalRecovered
- TotalCases, TotalTests
- TotalDeaths, TotalRecovered
- TotalDeaths, TotalTests
- TotalRecovered, ActiveCases
- TotalRecovered, TotalTests

Ta cũng thấy được biến dân số (Population) có ảnh hưởng trung bình (nằm ở mức 0.4) tới các biến TotalCases, TotalDeaths, TotalRecovered.

TotalCases ảnh hưởng vừa phải (0.58) đến ActiveCases.

Các cột đã chuẩn hóa (TotalCases/1M pop, ...) tương quan rất thấp với các cột chính (TotalCases, ...).

## 7. Các mối quan hệ trên 2 biến numerical

Nhóm thực hiện khảo sát giữa các cặp mối quan hệ sau:

- TotalDeaths ~ Population

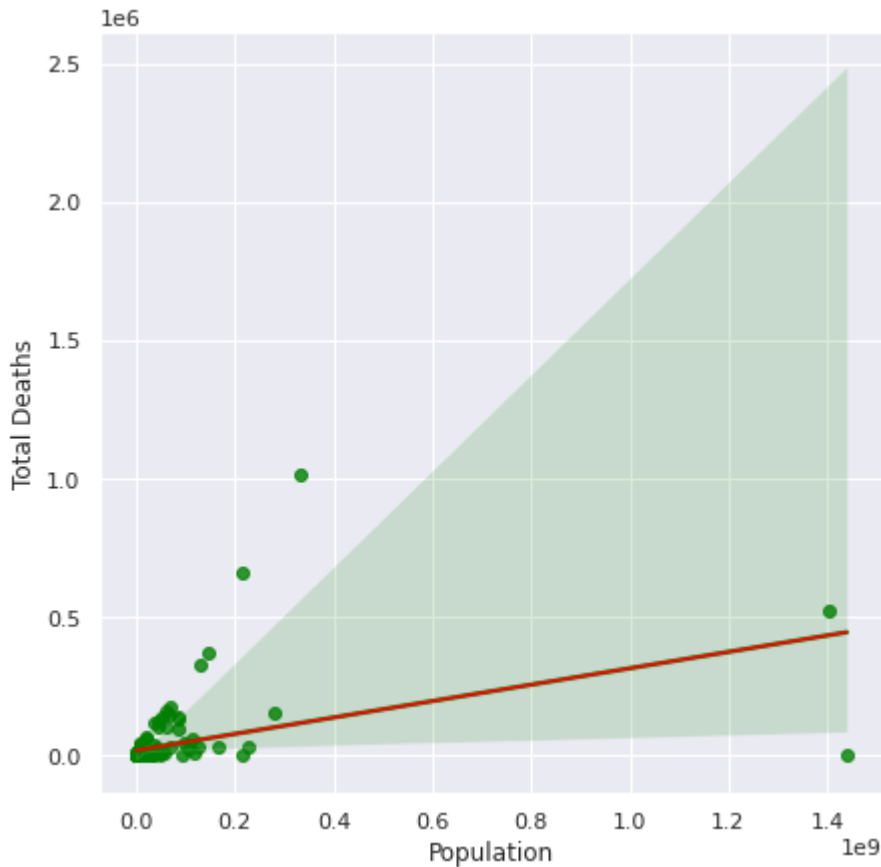
- TotalRecovered ~ TotalTests
- TotalDeaths ~ TotalCases
- ActiveCases ~ TotalCases.

#### a. TotalDeaths ~ Population

Mục đích: Nhóm lựa chọn cặp này để kiểm nghiệm xem “có phải dân số càng đông thì số ca tử vong càng nhiều không”

Kết quả khảo sát và biểu đồ trực quan:

OLS Regression Results						
<b>Dep. Variable:</b>	TotalDeaths	<b>R-squared:</b>	0.183			
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.179			
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	43.75			
<b>Date:</b>	Tue, 03 May 2022	<b>Prob (F-statistic):</b>	3.50e-10			
<b>Time:</b>	04:23:48	<b>Log-Likelihood:</b>	-2532.8			
<b>No. Observations:</b>	197	<b>AIC:</b>	5070.			
<b>Df Residuals:</b>	195	<b>BIC:</b>	5076.			
<b>Df Model:</b>	1					
<b>Covariance Type:</b> nonrobust						
	<b>coef</b>	<b>std err</b>	<b>t</b>	<b>P&gt; t </b>	<b>[0.025</b>	<b>0.975]</b>
<b>const</b>	1.796e+04	6864.551	2.617	0.010	4424.990	3.15e+04
<b>Population</b>	0.0003	4.49e-05	6.615	0.000	0.000	0.000
<b>Omnibus:</b>	267.778	<b>Durbin-Watson:</b>	1.888			
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	24558.133			
<b>Skew:</b>	5.720	<b>Prob(JB):</b>	0.00			
<b>Kurtosis:</b>	56.488	<b>Cond. No.</b>	1.58e+08			
<b>Warnings:</b>						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
[2] The condition number is large, 1.58e+08. This might indicate that there are strong multicollinearity or other numerical problems.						



Kết luận: Từ biểu đồ và kết quả khảo sát, ta thấy:

- Các điểm dữ liệu phân bố không đồng đều, chủ yếu hội tụ về góc dưới bên trái
- $R^2 = 0.183$ , chứng tỏ mô hình chưa thực sự tốt cho lắm
- Khoảng tin cậy “xòe” ra rất rộng về phía bên phải, chứng tỏ mô hình không thể dự đoán tốt mối quan hệ giữa 2 biến này.
- Biến Population đạt ý nghĩa thống kê ở mức 1% (do có  $p\text{-value} = 0.000 < 5\%$ ), mà hệ số của biến này  $> 0$  nên biến này có tương quan dương với biến TotalDeaths. Tuy nhiên, giá trị của hệ số này xấp xỉ 0, khẳng định cho bước tìm hiểu bằng heatmap của ta ở trên và trực quan khoảng tin cậy.

## b. TotalRecovered ~ TotalTests

Mục đích: Nhóm lựa chọn cặp này để kiểm nghiệm xem “có phải test càng nhiều thì càng có nhiều ca khỏi bệnh hay không”

Kết quả khảo sát và biểu đồ trực quan:



OLS Regression Results

Dep. Variable:

TotalRecovered

R-squared:

0.799

Model:

OLS

Adj. R-squared:

0.798

Method:

Least Squares

F-statistic:

774.8

Date:

Thu, 05 May 2022

Prob (F-statistic):

7.60e-70

Time:

03:13:49

Log-Likelihood:

-3241.1

No. Observations:

197

AIC:

6486.

Df Residuals:

195

BIC:

6493.

Df Model:

1

Covariance Type:

nonrobust

coef

std err

t

P>|t|

[0.025

0.975]

const

3.891e+05

2.51e+05

1.548

0.123

-1.07e+05

8.85e+05

TotalTests

0.0612

0.002

27.835

0.000

0.057

0.066

Omnibus:

139.941

Durbin-Watson:

2.323

Prob(Omnibus):

0.000

Jarque-Bera (JB):

5066.454

Skew:

2.086

Prob(JB):

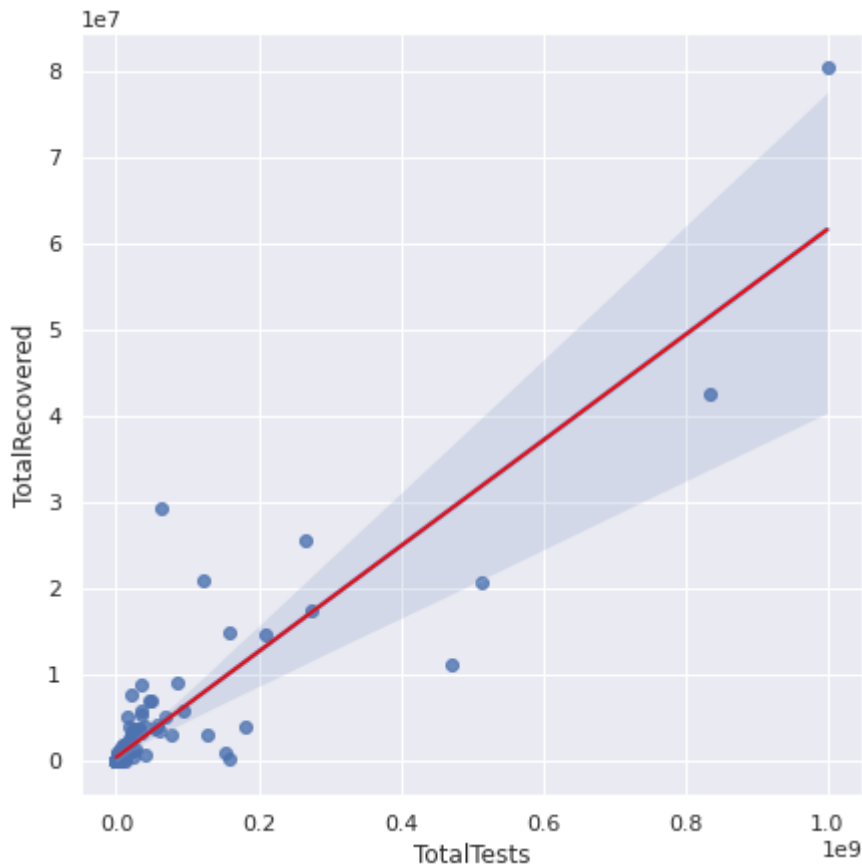
0.00

Kurtosis:

27.491

Cond. No.

1.19e+08



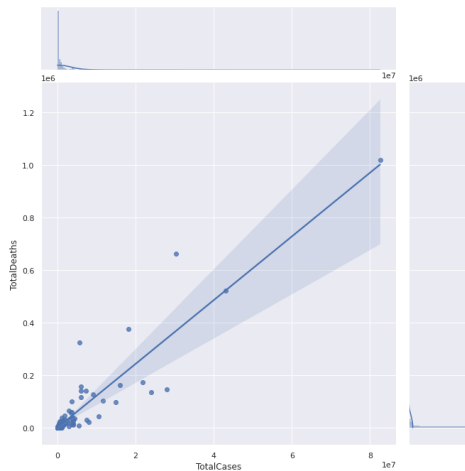
Kết luận: Từ biểu đồ và kết quả khảo sát, ta thấy:

- Các điểm dữ liệu phân bố không đồng đều, chủ yếu hội tụ về góc dưới bên trái
- $R^2 = 0.799$ , chứng tỏ mô hình khá tốt.
- Biến TotalRecovered đạt ý nghĩa thống kê ở mức 1% (do có p-value = 0.000 < 5%), mà hệ số của biến này > 0 nên biến này có tương quan dương với biến TotalTests

### c. TotalDeaths ~ TotalCases

Mục đích: Nhóm lựa chọn cặp này để kiểm nghiệm xem “có phải càng nhiễm nhiều thì số ca tử vong có nhiều theo hay không”

Kết quả khảo sát và biểu đồ trực quan:



OLS Regression Results					
Dep. Variable:	TotalDeaths	R-squared:	0.852		
Model:	OLS	Adj. R-squared:	0.852		
Method:	Least Squares	F-statistic:	1126.		
Date:	Tue, 03 May 2022	Prob (F-statistic):	6.01e-83		
Time:	09:29:54	Log-Likelihood:	-2364.3		
No. Observations:	197	AIC:	4733.		
Df Residuals:	195	BIC:	4739.		
Df Model:	1				
Covariance Type: nonrobust					
	coef	std err	t	P> t	[0.025 0.975]
const	205.4686	2955.286	0.070	0.945	-5622.959 6033.896
TotalCases	0.0121	0.000	33.557	0.000	0.011 0.013
Omnibus:	165.090	Durbin-Watson:	1.897		
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6568.603		
Skew:	2.691	Prob(JB):	0.00		
Kurtosis:	30.772	Cond. No.	8.56e+06		
Warnings:					
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.					
[2] The condition number is large, 8.56e+06. This might indicate that there are strong multicollinearity or other numerical problems.					

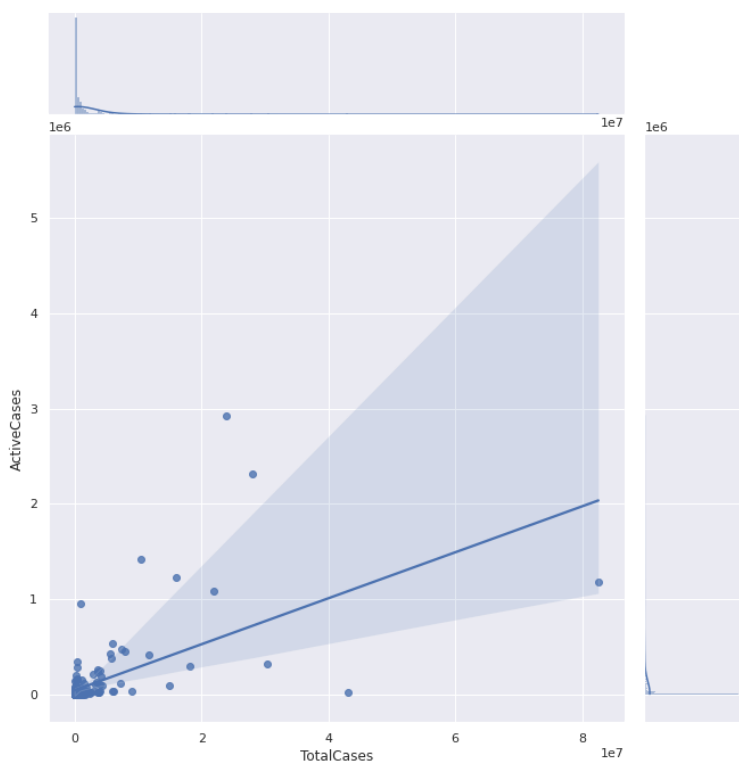
Kết luận: Từ biểu đồ và kết quả khảo sát, ta thấy:

- Các điểm dữ liệu phân bố không đồng đều, chủ yếu hội tụ về góc dưới bên trái
- $R^2 = 0.852$ , chứng tỏ mô hình này khá tốt
- Biến TotalCases đạt ý nghĩa thống kê ở mức 1% (do có p-value = 0.000 < 5%), mà hệ số của biến này > 0 nên biến này có tương quan dương với biến TotalDeaths

### d. ActiveCases ~ TotalCases

Mục đích: Nhóm lựa chọn cặp này để kiểm nghiệm xem “có phải tổng ca nhiễm nhiều thì hiện tại có nhiều ca đang dương tính hay không?”

OLS Regression Results						
Dep. Variable:	ActiveCases	R-squared:	0.332			
Model:	OLS	Adj. R-squared:	0.329			
Method:	Least Squares	F-statistic:	97.05			
Date:	Thu, 05 May 2022	Prob (F-statistic):	7.71e-19			
Time:	03:34:34	Log-Likelihood:	-2741.4			
No. Observations:	197	AIC:	5487.			
Df Residuals:	195	BIC:	5493.			
Df Model:	1					
Covariance Type: nonrobust						
	coef	std err	t	P> t	[0.025	0.975]
const	4.118e+04	2e+04	2.055	0.041	1651.819	8.07e+04
TotalCases	0.0241	0.002	9.852	0.000	0.019	0.029
Omnibus:	228.951	Durbin-Watson:	1.297			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11404.539			
Skew:	4.570	Prob(JB):	0.00			
Kurtosis:	39.137	Cond. No.	8.56e+06			



Kết luận: Từ biểu đồ và kết quả khảo sát, ta thấy:

- Các điểm dữ liệu phân bố không đồng đều, chủ yếu hội tụ về góc dưới bên trái.
- $R^2 = 0.332$  và khoảng tin cậy rất rộng ở phía bên phải, chứng tỏ mô hình này chưa tốt.
- Biến *TotalCases* đạt ý nghĩa thống kê ở mức 1% (do có  $p\text{-value} = 0.000 < 5\%$ ), mà hệ số của biến này  $> 0$  nên biến này có tương quan dương đến biến *ActiveCases*.

## 8. Mối quan hệ giữa các cột *NewCases*, *NewDeaths*, *NewRecovered* sau khi xử lý giá trị null

Khi loại bỏ giá trị null ở 3 cột này đi bộ dữ liệu chỉ còn có 58 dòng, số lượng khá ít so với 228 dòng ban đầu, nên nhóm sẽ xử lý các giá trị null như sau

- Nguyên nhân dẫn tới các giá trị bị null:
  - + Do không có ca nhiễm/tử vong/ khỏi bệnh tăng
  - + Do hệ thống chưa cập nhật được dữ liệu từ các quốc gia (các dòng có giá trị null ở cột ở các cột *NewCases*, *NewDeaths*, *NewRecovered*)
- Giải pháp: nhóm sẽ lấy dữ liệu các cột *TotalCases*, *NewCases*, *TotalDeaths* của ngày 21/4 trừ cho các cột *TotalCases*, *NewCases*, *TotalDeaths* của ngày 20/4 sẽ ra được các cột *NewCases*, *NewDeaths*, *NewRecovered* của ngày 21/4

Mục đích: kiểm chứng liệu khi số ca nhiễm phát hiện trong ngày tăng thì số ca tử vong, khỏi bệnh trong ngày có tăng theo không? số ca hồi phục trong ngày tăng thì số ca tử vong có tăng theo không

NewCases ~ NewRecovered

---

OLS Regression Results

<b>Dep. Variable:</b>	NewCases	<b>R-squared:</b>	0.758
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.756
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	634.7
<b>Date:</b>	Tue, 03 May 2022	<b>Prob (F-statistic):</b>	2.11e-64
<b>Time:</b>	13:37:07	<b>Log-Likelihood:</b>	-2115.8
<b>No. Observations:</b>	205	<b>AIC:</b>	4236.
<b>Df Residuals:</b>	203	<b>BIC:</b>	4242.
<b>Df Model:</b>	1		

**Covariance Type:** nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	787.1792	526.670	1.495	0.137	-251.266	1825.624
NewRecovered	0.5885	0.023	25.193	0.000	0.542	0.635

**Omnibus:** 200.844 **Durbin-Watson:** 2.197  
**Prob(Omnibus):** 0.000 **Jarque-Bera (JB):** 8375.666  
**Skew:** 3.516 **Prob(JB):** 0.00  
**Kurtosis:** 33.514 **Cond. No.** 2.30e+04

NewCases ~ NewDeaths

OLS Regression Results

<b>Dep. Variable:</b>	NewCases	<b>R-squared:</b>	0.418
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.415
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	145.6
<b>Date:</b>	Tue, 03 May 2022	<b>Prob (F-statistic):</b>	1.27e-25
<b>Time:</b>	13:37:08	<b>Log-Likelihood:</b>	-2205.6
<b>No. Observations:</b>	205	<b>AIC:</b>	4415.
<b>Df Residuals:</b>	203	<b>BIC:</b>	4422.
<b>Df Model:</b>	1		

**Covariance Type:** nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	1156.4540	822.461	1.406	0.161	-465.208	2778.116
NewDeaths	154.5489	12.809	12.065	0.000	129.293	179.805

**Omnibus:** 153.404 **Durbin-Watson:** 1.770  
**Prob(Omnibus):** 0.000 **Jarque-Bera (JB):** 9705.885  
**Skew:** 2.091 **Prob(JB):** 0.00  
**Kurtosis:** 36.449 **Cond. No.** 66.1

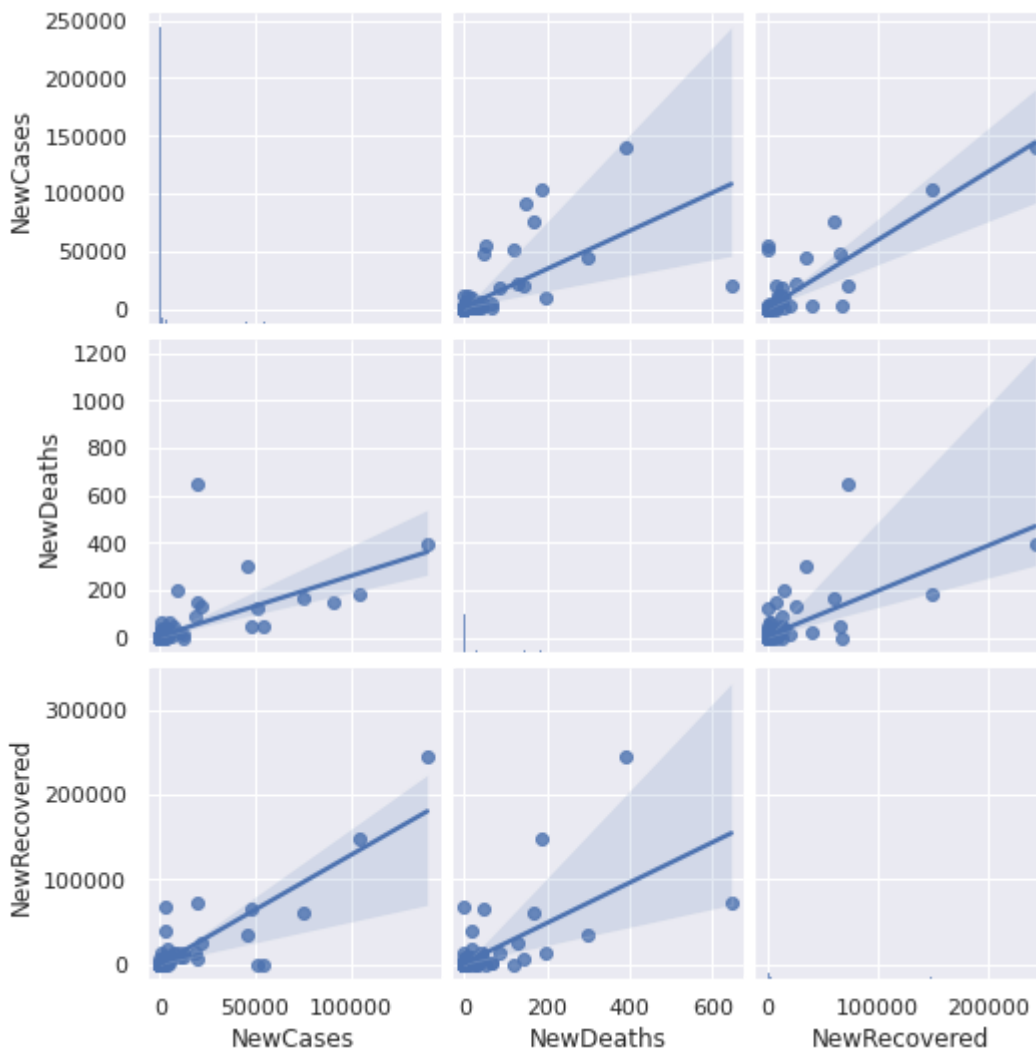
NewRecovered ~ NewDeaths

# OLS Regression Results

**Dep. Variable:** NewRecovered **R-squared:** 0.451  
**Model:** OLS **Adj. R-squared:** 0.448  
**Method:** Least Squares **F-statistic:** 166.6  
**Date:** Tue, 03 May 2022 **Prob (F-statistic):** 3.21e-28  
**Time:** 13:38:04 **Log-Likelihood:** -2279.9  
**No. Observations:** 205 **AIC:** 4564.  
**Df Residuals:** 203 **BIC:** 4570.  
**Df Model:** 1  
**Covariance Type:** nonrobust

	coef	std err	t	P> t	[0.025	0.975]
const	1008.2956	1181.369	0.853	0.394	-1321.032	3337.624
NewDeaths	237.4724	18.399	12.907	0.000	201.195	273.750

**Omnibus:** 247.347 **Durbin-Watson:** 1.956  
**Prob(Omnibus):** 0.000 **Jarque-Bera (JB):** 17943.726  
**Skew:** 4.753 **Prob(JB):** 0.00  
**Kurtosis:** 47.837 **Cond. No.** 66.1



- Từ việc thống kê về các mối quan hệ, ta thấy được  $R^2$  khá cao, lần lượt là 0.758, 0.418, 0.451 cùng với đó p-value xấp xỉ gần bằng 0, hệ số các biến dương nên các biến này có tương quan dương với nhau
- Các quốc gia có số ca nhiễm ghi nhận hàng ngày càng nhiều thì số ca khỏi bệnh hàng ngày cũng nhiều
- Dù đã có vaccine nhưng một số quốc gia có số ca nhiễm ghi nhận hàng ngày càng nhiều thì số ca tử vong hàng ngày cũng nhiều
- Các quốc gia có số ca khỏi bệnh ghi nhận hàng ngày càng nhiều thì số ca tử vong hàng ngày cũng nhiều.

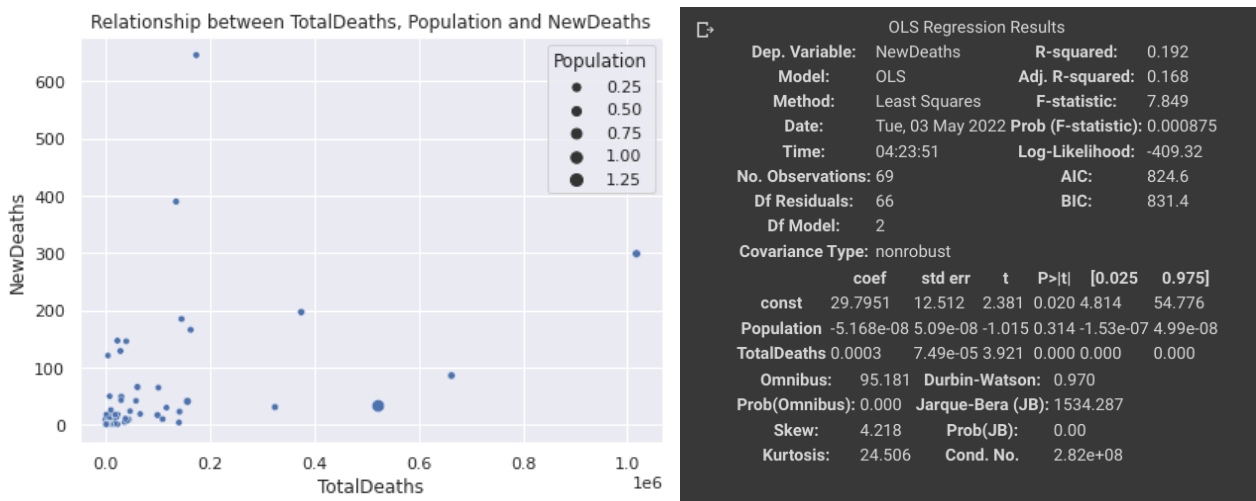
## 9. Các mối quan hệ trên nhiều hơn 2 biến numerical

Nhóm sẽ lần lượt xem xét các mối quan hệ sau:

- $\text{NewDeaths} \sim \text{TotalDeaths} + \text{Population}$
- $\text{TotalDeaths} \sim \text{TotalCases} + \text{TotalTests}$
- $\text{NewCases} \sim \text{NewRecovered} + \text{ActiveCases}$

### a. $\text{NewDeaths} \sim \text{TotalDeaths} + \text{Population}$

Ở đây nhóm sẽ phân tích liệu rằng *có phải các quốc gia có dân số đông, và hiện đang có nhiều ca nhiễm thì cũng sẽ tiếp tục có nhiều ca nhiễm hay không.*



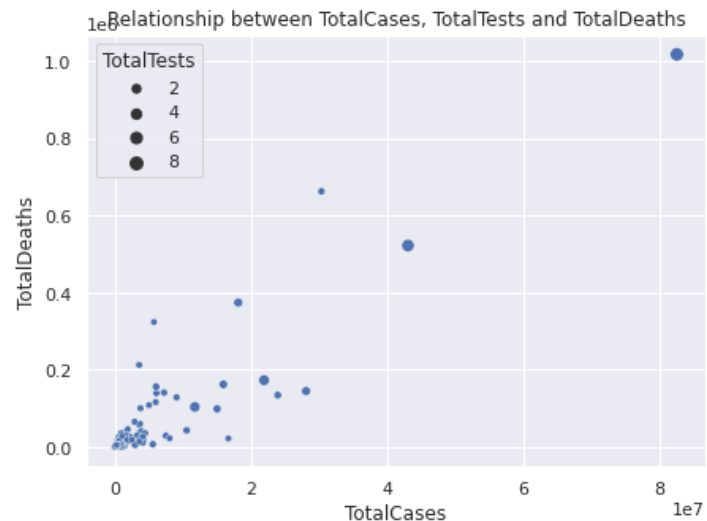
Thông qua trực quan hóa, ta nhận thấy đúng là số lượng ca hiện tại có lẽ có ảnh hưởng đến số lượng ca mới trong ngày 21/4. Tuy nhiên cũng có rất nhiều trường hợp tổng số ca nhiễm rất lớn nhưng số ca ghi nhận là rất thấp, hoặc ngược lại. Đây có thể là các outlier, và chúng có ảnh hưởng khá lớn đến các model mà ta đã xây dựng.

$R^2$  của mô hình là rất thấp (0.192) và p-value của biến Population khá lớn ( $0.3 > 0.05$ ) nên nó không có ý nghĩa thống kê. Hệ số của các biến độc lập cũng xấp xỉ 0.

### b. TotalDeaths ~ TotalCases + TotalTests

Ở đây nhóm sẽ phân tích liệu rằng *có phải các quốc gia có nhiều ca nhiễm, và hiện đã test diện rộng rồi thì cũng sẽ có nhiều ca tử vong hay không.*

OLS Regression Results					
Dep. Variable:	TotalDeaths	R-squared:	0.828		
Model:	OLS	Adj. R-squared:	0.827		
Method:	Least Squares	F-statistic:	496.7		
Date:	Thu, 05 May 2022	Prob (F-statistic):	1.56e-79		
Time:	03:41:42	Log-Likelihood:	-2520.2		
No. Observations:	209	AIC:	5046.		
Df Residuals:	206	BIC:	5056.		
Df Model:	2				
Covariance Type: nonrobust					
	coef	std err	t	P> t	[0.025 0.975]
const	845.9264	3050.917	0.277	0.782	-5169.100 6860.953
TotalTests	-0.0001	5.74e-05	-2.004	0.046	-0.000 -1.85e-06
TotalCases	0.0134	0.001	16.656	0.000	0.012 0.015
Omnibus:	110.100	Durbin-Watson:	1.880		
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3524.974		
Skew:	1.338	Prob(JB):	0.00		
Kurtosis:	22.940	Cond. No.	1.17e+08		



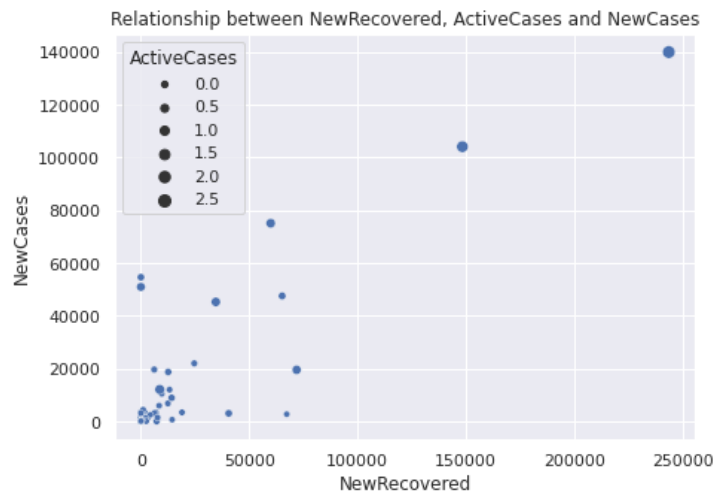
Nhìn vào trực quan, ta thấy được: tổng số ca càng cao thì số ca tử vong cũng càng cao, và dân số càng đông thì tổng số ca tử vong cũng cao (nhìn vào kích thước của các chấm tròn).

Nhìn vào kết quả hồi quy, ta thấy mô hình là khá tốt với  $R^2$  ở mức 0.82. Tuy nhiên biến TotalTests có vẻ đóng góp không nhiều vào kết quả hồi quy, nên cơ bản đây chỉ là mô hình hồi quy giữa 2 biến mặc dù tương quan giữa TotalTests và TotalDeaths là khá cao.

### c. NewCases ~ NewRecovered + ActiveCases

Ở đây nhóm sẽ phân tích liệu rằng *có phải các quốc gia hiện đang có nhiều ca dương tính, nhưng hôm nay có nhiều ca khỏi bệnh thì sẽ tiếp tục có thêm nhiều ca dương tính hay không.*

OLS Regression Results					
Dep. Variable:	NewCases	R-squared:	0.846		
Model:	OLS	Adj. R-squared:	0.845		
Method:	Least Squares	F-statistic:	577.5		
Date:	Thu, 05 May 2022	Prob (F-statistic):	4.39e-86		
Time:	03:56:57	Log-Likelihood:	-2146.1		
No. Observations:	213	AIC:	4298.		
Df Residuals:	210	BIC:	4308.		
Df Model:	2				
Covariance Type: nonrobust					
	coef	std err	t	P> t	[0.025 0.975]
const	-220.4192	414.528	-0.532	0.595	-1037.588 596.750
NewRecovered	0.2531	0.036	7.095	0.000	0.183 0.323
ActiveCases	0.0268	0.002	10.966	0.000	0.022 0.032
Omnibus:	132.463	Durbin-Watson:	2.056		
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4904.383		
Skew:	1.718	Prob(JB):	0.00		
Kurtosis:	26.255	Cond. No.	3.44e+05		



Nhìn vào trực quan, ta thấy số ca khỏi bệnh hôm nay càng cao thì số ca nhiễm mới cũng càng cao, cho thấy tình hình dịch vẫn đang “giằng co”. Ngoài ra, những nước có tổng số ca nhiễm càng cao thì hôm nay cũng có nhiều ca nhiễm mới hơn các nước khác.

Về mặt thống kê, mô hình cho kết quả khá tốt với  $R^2$  đạt 0.84, các biến độc lập đều có ý nghĩa thống kê với p-value bằng 0 và hệ số dương.

## V. Tài liệu tham khảo

- Slide lý thuyết của môn học (thầy Bùi Tiến Lên).
- [Pandas documentation](#)
- [Linear Regression, Statsmodel documentation](#)
- [Seaborn documentation](#)
- [Matplotlib documentation](#)
- [BeautifulSoup documentation](#)